DOCUMENT RESUME

ED 306 265 TM 013 075

AUTHOR Woodruff, David
TITLE Derivation and Application of the Conditional Standard Error of Measurement in Prediction.
PUB DATE Mar 89
NOTE 38p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Error of Measurement; *Estimation (Mathematics); *Mathematical Models; *Predictive Measurement
IDENTIFIERS *Conditionals

ABSTRACT

Previous methods for estimating the conditional standard error of measurement (CSEM) at specific score or ability levels are critically discussed, and a brief summary of prior empirical results is given. A new method is developed which avoids theoretical problems inherent in some prior methods, is easy to implement, and estimates not only a quantity analogous to the CSEM at each score but also the conditional standard error of prediction (CSEP) at each score and the conditional reliability at each score. The new method and several older methods are applied to a variety of tests, and representative results are illustrated by graphical display. The CSEM-like estimates produced by the new method are called conditional standard error of measurement in prediction (CSEMP) estimates and are similar to those produced by older methods, but the CSEP estimates produced by the new method offer an alternative interpretation of the accuracy of a test at different scores. Finally, evidence is presented which shows that previous methods can produce dissimilar results, and that the shape of the score distribution may influence the way in which the CSEM varies across the score scale. (Fourteen graphs are included.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED306265

Derivation and Application of the Conditional Standard Error of Measurement in Prediction

David Woodruff
The American College Testing Program
P.O. Box 168
Iowa City, Iowa 52243

Paper presented at the March, 1989 Annual Meeting of
The American Educational Research Association
San Francisco

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

DAVID J. WOODRUFF

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Acknowledgement: The author benefited from discussions with Leonard Feldt during the development of this article.

Running Head: ERROR OF MEASUREMENT

TM013075

Abstract

Previous methods for estimating the conditional standard error of measurement (CSEM) at specific score or ability levels are critically discussed, and a brief summary of prior empirical results is given. A new method is developed which avoids theoretical problems inherent in some prior methods, is easy to implement, and estimates not only a quantity analogous to the CSEM at each score but also the conditional standard error of prediction (CSEP) at each score and the conditional reliability at each score. The new method and several older methods are applied to a variety of tests, and representative results are illustrated by graphical display. The CSEM-like estimates produced by the new method are called conditional standard error of measurement in prediction (CSEMP) estimates and are similar to those produced by older methods, but the CSEP estimates produced by the new method offer an alternative interpretation of the accuracy of a test at different scores. Finally, evidence is presented which shows that previous methods can produce dissimilar results, and that the shape of the score distribution may influence the way in which the CSEM varies across the score scale.

Key words: ANOVA, standard error of measurement, regression, reliability, test theory

Introduction

Several recent articles have addressed the issue of estimating conditional standard errors of measurement (CSEM) for given ability levels or observed scores: Livingston (1982); Lord (1984); Feldt, Steffen, and Gupta (1985); Blixt and Shama (1986); and Jarjoura (1986). Other more distant references are Mollenkopf (1949), Thorndike (1951), Lord (1955), and Keats (1957). Lord (1984) suggests that the current interest in this topic stems from the increased use of cutting scores on criterion referenced tests and the recommendation contained in the Standards for Educational and Psychological Testing (APA, 1985) that such estimates be provided for a variety of tests. Whatever the reasons for this resurgence in interest, the above references present theoretical and empirical evidence which indicates that CSEM values differ at different ability or score levels and, therefore, that the average standard error of measurement (SEM) may be an unsatisfactory index in certain circumstances. For example, test developers may wish to control the maximum CSEM or the CSEM at one or more crucial ability or score levels rather than the SEM, while test users may find CSEM values useful in interpreting individual scores.

Several methods are available for estimating CSEM. These may be divided into two classes: those based on strong true score models and those based on weak true score models (Lord and Novick, 1968). In the former class are methods derived from the binomial error model (Lord, 1955, 1984 and Feldt et al., 1985) and from parametric item response models (Lord, 1984 and Feldt et al., 1985). The latter class include what will be called here difference methods (Mollenkopf, 1949; Thorndike, 1951; Livingston, 1982; Lord, 1984; and Feldt et al., 1985) and ANOVA methods (Feldt et al., 1985 and Jarjoura, 1986). The difference methods begin by dividing a test into two parallel

half-tests and then take their difference; some use regression techniques to smooth the estimates. The ANOVA methods derive from Hoyt's (1941) demonstration that reliability and error variance may be estimated from a two-way items by examinees ANOVA.

The paper begins with a critical evaluation of current methods for estimating CSEM. Part of this discussion is an elaboration of criticisms voiced by Lord (1984), and it motivates the subsequent derivation of an alternative method which estimates a quantity analogous to CSEM, and that is labeled conditional standard error of measurement in prediction (CSEMP). This new method derives from Hoyt's (1941) results and is related to the methods of Feldt et al. (1985) and Jarjoura (1986).

Three of the recent investigations (Lord, 1984; Feldt et al., 1985; and Blixt and Shama, 1986) have concluded that the various methods yield essentially the same results. The estimated CSEM's have a maximum at the score scale midpoint and decrease symmetrically to their minimum at the score scale end points as the binomial error model predicts. However, the authors warn that their results are specific to their data and that differently designed tests may yield different results. Jarjoura (1986) found that his ANOVA method and the difference method reached a maximum below the score scale midpoint and decreased asymmetrically. Jarjoura studied a logical reasoning test with a negatively skewed score distribution and a mean proportion correct score of .65. The other authors studied achievement or aptitude tests of either a verbal or quantitative nature with mean proportion correct scores closer to .50 and more symmetrical score distributions. The sample sizes in these studies ranged from 2,000 to 17,000 examinees.

The paper concludes with an application of several methods including the new one to four achievement tests with sample sizes of about 40,000 and two

forms of a licensure examination both with sample sizes of approximately 4,000. The four achievement tests have mean proportion-correct scores near .50 and fairly symmetric score distributions. The licensure examinations have negatively skewed score distributions with mean proportion-correct scores near .80. The results are illustrated by graphical displays.

Critique of Previous Methods

Lord (1984) and Feldt et al. (1985) illustrate the application of IRT models in obtaining CSEM estimates. Two criticisms of IRT models are that they require restrictive assumptions and that item parameter estimation is complex both theoretically and practically (Baker, 1987). After item parameter estimates are obtained, CSEM estimates can be computed for any value of the true score ξ. Such estimates can be used to compare different test forms in test construction or to compare different tests in test selection. In applying the method to a group of examinees, the true scores of the examinees are not known, and CSEM values are assigned according to estimated true score, $\hat{\xi}$, instead. This is a reasonable approach under IRT models, but as shortly will be discussed, conditioning on observed scores is different from conditioning on true scores, and at least to the extent that $\rho(X,\hat{\xi})$ exceeds the square root of the reliability of the test, $\rho(X,\xi)$, conditioning on estimated true scores is similar to conditioning on observed scores.

Lord (1984) and Feldt et al. (1985) also illustrate the application of the binomial error model as well as several variations on it in estimating CSEM. Under the binomial error model, the formula for the squared CSEM of examinees with true score ξ is $\xi(m - \xi)/m$ where m is the number of items in the test. The binomial error model requires that all dichotomously scored m item tests have the same CSEM values. This is a strong assumption. Lord (1984) notes that if each examinee is given a different random sample of m

items, then an unbiased estimate of the squared CSEM for a randomly selected examinee with observed score X is $X(m - X)/(m - 1)$. Lord (1984) further notes that it is not appropriate to assign the above CSEM estimate to all examinees with observed score X in a group of examinees who have taken the same m items because the conditioning is then on the observed score and not on the true score. Van der Linden (1979) discusses in detail assumptions made by the binomial error model.

Let $X = T + E$ denote the m item full test and $Y_1 = T/2 + E_1$ and $Y_2 = T/2 + E_2$ the classically parallel half-tests into which X is divisible. Difference methods use $s^2(Y_1 - Y_2|X = q) = s^2(E_1 - E_2|X = q)$ as an estimate of the CSEM, $\sigma^2(X|T = q) = \sigma^2(E|T = q)$. Under the usual assumptions of classical test theory, the errors, E_1 and E_2, are uncorrelated over repeated measurements on the same examinee, and the errors are also uncorrelated in a subsample of examinees all of whom have the same true score. Contrarily, E_1 and E_2 are negatively correlated within a subsample of examinees having the same observed score. The correlation arises because the conditioning is on X and not on T. For the subsample of examinees all with the same observed score, $X = T + E_1 + E_2$ is constant and this linear restriction induces pairwise negative correlations between the three random variables, T, E_1, and E_2 (Aitchinson, 1986). As a consequence, difference methods are actually estimating

$$\sigma^2(E_1 - E_2|X = q) = \sigma^2(E_1|X = q) + \sigma^2(E_2|X = q) - 2\sigma(E_1,E_2|X = q), \quad (1)$$

which does not equal $\sigma^2(E|X = q)$ because the covariance term is being subtracted and not added.

It is shown in the appendix that

$$\mu[\sigma^2(E|X)] \leq \mu[\sigma^2(E|T)] = \sigma^2(E) \quad (2)$$

with equality holding if and only if $\mu(T|X) = X$ for all X, which is not true in general. It is also shown that if the regression of T on X is linear, which may occur when E and T have marginal distributions similar in shape (Lindley, 1947), then

$$\mu[\sigma^2(E|X)] = \rho(X,X')\sigma^2(E) = \rho(X,X')\mu[\sigma^2(E|T)] \ . \qquad (3)$$

Equation (1) shows that on average the difference method is overestimating $\sigma^2(E|X = q)$ because of the negative conditional covariance between E_1 and E_2, but this may make the difference method estimate a better estimate of $\sigma^2(E|T = q)$, at least on average, because of (2) and (3). Nonetheless, the difference method is conditioning on the wrong quantity and at best is only providing approximate estimates for $\sigma^2(E|T)$.

The ANOVA method of Feldt et al. (1985) estimates $\sigma^2(E|T = q)$ by computing $m^*(MS_{PM}|X = q)$ where this conditional interaction mean square is computed from a two-way persons (examinees) by measures (items) ANOVA on the subsample of examinees having observed scores equal to q. However, this estimate is more appropriately considered as an estimate of $\sigma^2(E|X = q)$ which for comparison to (1) may be expressed as:

$$\begin{aligned}\sigma^2(E|X = q) &= \sigma^2(E_1 + E_2|X = q) \\ &= \sigma^2(E_1|X = q) + \sigma^2(E_2|X = q) + 2\sigma(E_1,E_2|X = q) \qquad (4)\end{aligned}$$

and which should always be less than $\sigma^2(E_1 - E_2|X = q)$ in (1) due to the negative conditional covariance between E_1 and E_2. To the extent to which this ANOVA method is accurately estimating $\sigma^2(E|X = q)$, equations (2) and (3) show that this ANOVA method may be underestimating $\sigma^2(E|T)$, at least on average, but as the reliability of the test increases the average of this underestimation decreases.

The ANOVA method of Jarjoura (1986) allows for the calculation of the CSEM for a specific examinee and as such is immune from the foregoing

criticism of conditioning on the wrong quantity. If, however, the estimated CSEM's for all examinees with a given score are averaged to obtain an estimate of the CSEM at that score, as Jarjoura (1986) does, then this estimate, which is similar but not identical to the ANOVA estimate of Feldt et al. (1985), is also subject to the criticism that it is conditioning on X and not on T.

A final criticism of estimating $\sigma^2(E|X = q)$ in place of $\sigma^2(E|T = q)$ is that $\sigma^2(E|X = q)$ is constrained to equal $\sigma^2(T|X = q)$ for all values of q. To see this, note that because $X=T+E$, $\sigma^2(E|X=q) = \sigma^2(X-T|X=q) = \sigma^2(q-T|X=q) = \sigma^2(T|X=q)$ since q is constant.

In the next section, an alternative ANOVA method is developed which retains the simplicity of the weak true score methods just discussed while avoiding their theoretical shortcomings.

The Alternative Method

Consider two classically parallel full-length tests each with m items, $X = T + E$ and $X' = T + E'$, both of which have been administered to N examinees. Since estimating $\sigma^2(X|T)$ is equivocal when T is not observed, the alternative method considers

$$\sigma^2(X'|X) = \sigma^2(T + E'|X) = \sigma^2(T|X) + \sigma^2(E'|X)$$

where T and E′ are uncorrelated given $X = T + E$ because T, E, and E′ are mutually uncorrelated for each examinee by assumption. Since it is also assumed that over repeated measures on the same examinee $\mu(E) = \mu(E') = 0$ for each and every examinee, it follows that $\mu(E'|X) = 0$ for all X and therefore that $\mu[\sigma^2(E'|X)] = \sigma^2(E')$. As a consequence, if $\sigma^2(E'|X) = 0$ for all X, then $\sigma^2(E') = \sigma^2(E) = 0$ and $X' = X = T$. This is not true for $\sigma^2(E|X)$.

Within subgroups defined by conditioning on $X = q$ for various values of q, a persons (examinees) by measures (items) ANOVA as illustrated by Hoyt

(1941) can be executed on X′ to obtain estimates for the following parameters: $\sigma^2(X'|X = q)$, $\sigma^2(T|X = q)$, $\sigma^2(E'|X = q)$, and $\alpha(X'|X = q) = \sigma^2(T|X = q)/\sigma^2(X'|X = q)$, this last parameter being a conditional coefficient alpha. What is being done is that the residual observed score variances obtained from regressing X′ onto X are being decomposed into true score and error score variances. Lord and Novick (1968, chap. 3) refer to the average residual observed score variance obtained from regressing X′ onto X as the standard error of prediction (SEP), which suggests conditional standard error of measurement in prediction (CSEMP) as the name for that part of the conditional standard errors of prediction (CSEP) which is due only to measurement error variance and not to true score variance.

The above conditional parameters are easily interpreted as describing the performance of examinees on retesting given their scores on a previous parallel test, and as such offer information about how accurately the examination predicts examinee performance at different score levels. The use and interpretation of these conditional parameters will be further discussed in the context of the examples presented later. A practical problem with obtaining estimates for these parameters is that data on two parallel tests is rarely available. The remainder of this section is devoted to obtaining such estimates from a single test when the test is divisible into classically parallel half-tests, though it should also be clear how to proceed when scores on two full-length tests are available. Strong distributional assumptions are not required.

Again, let the m item full length test $X = T + E$ be divisible into two parallel half-tests, $Y_1 = T/2 + E_1$ with m_1 items and $Y_2 = T/2 + E_2$ with m_2 items, and let the linear transformation $X^* = bY_1 + c$ rescale Y_1 so that X^* has the same mean and variance as X, i.e., X^* is Y_1 linearly equated to

X. (Sometimes it is assumed that two tests must be equally reliable to be so equated, but that assumption is not necessary in this application.) The method begins by dividing the N examinees into subgroups according to their X^* scores and then applying a persons (P) by measures (M) ANOVA to the item responses on Y_2 in each subgroup. This leads to the following estimates for the subgroups $X^* = q$:

$$s^2(Y_2|X^*=q) = m_2(MS_P|X^*=q), \tag{5}$$

$$s^2(E_2|X^*=q) = m_2(MS_{PM}|X^*=q), \tag{6}$$

$$s^2(T/2|X^*=q) = s^2(Y_2|X^*=q) - s^2(E_2|X^*=q) \tag{7}$$

$$= m_2[(MS_P|X^*=q) - (MS_{PM}|X^*=q)], \quad \text{and}$$

$$a(Y_2|X^*=q) = s^2(T/2|X^*=q)/s^2(Y_2|X^*=q). \tag{8}$$

The above statistics are for the conditional distributions of Y_2 given X^*, but what is needed are estimates for the conditional distributions of X', a parallel measure of X, given X. Such estimates may be obtained by applying the "step-up" formulas for composite measures (Lord and Novick, 1968, chap. 4) to the Y_2 given X^* estimates. This results in

$$s^2(X'|X^*=q) = s^2(T|X^*=q) + s^2(E'|X^*=q) \quad \text{where} \tag{9}$$

$$s^2(E'|X^*=q) = 2s^2(E_2|X^*=q) = 2m_2(MS_{PM}|X^*=q), \tag{10}$$

$$s^2(T|X^*=q) = 4s^2(T/2|X^*=q) = 4m_2[(MS_P|X^*=q) - (MS_{PM}|X^*=q)], \quad \text{and} \tag{11}$$

$$a(X'|X^*=q) = \frac{2a(Y_2|X^*=q)}{1+a(Y_2|X^*=q)} = \frac{s^2(T|X^*=q)}{s^2(X'|X^*=q)} . \tag{12}$$

The method assumes that conditioning on X^* approximates conditioning on X, and that the "step-up" formulas of Lord and Novick (1968, chap. 4), which in this application state that doubleing the test length doubles the error score variance but quadruples the true-score variance, hold conditionally and not just marginally. These assumptions are not easily testable unless a full

length parallel test is available, in which case they are not needed. In the next section, results from applying the method to several tests are presented, and these results suggest that the above assumptions yield reasonable estimates.

Two Examples

Several methods for estimating CSEM including the one just developed were applied to the four subtests of the ACT Assessment which are English (75 items), Mathematics (40 items), Social Studies (52 items) and Natural Sciences (52 items) and to two forms of a professional licensure examination (300 items). Data for a random sample of 40,000 examinees were available for the ACT Assessment subtests, and each of the two forms of the licensure examination had been administered to approximately 4,000 examinees. The ACT Assessment subtests were divided by odd-even splits into parallel half-tests. The licensure examinations were composed of separately timed morning and afternoon sections which were parallel by construction. Score distributions for the four ACT Assessment subtests were generally symmetric and platykurtic with means at the score scale midpoints. The licensure examinations had negatively skewed score distributions with mean proportion-correct scores near .80.

Results for the four ACT Assessment subtests were very similar and so results for only one subtest, Natural Sciences, are presented. Figure 1 presents unsmoothed graphs of CSEM estimates derived from the binomial method, the ANOVA method of Feldt et al. (ANOVA1) and the present ANOVA method (ANOVA2) whose estimates are actually of CSEMP. Both ANOVA methods used grouped data with an interval width of 4 score points, whereas the binomial estimates are derived from ungrouped data though only every other score point is plotted. The binomial estimates exceed the other estimates over most of

the score scale. Because the binomial model implies KR21 as a reliability estimate, Lord (1984) considers it to yield overestimates of CSEM when the same items are administered to all examinees. After converging around the center of the score scale, the two ANOVA methods diverge at both extremes with the ANOVA2 estimates exceeding the ANOVA1 estimates. The ANOVA1 estimates decrease symmetrically from their maximum near the score scale midpoint, as do the binomial estimates, but the ANOVA2 estimates decrease asymmetrically. The ANOVA2 estimates of CSEMP are more homogeneous over the score scale than are the other methods' estimates of CSEM. This is due to the fact that the ANOVA2 method conditions on a parallel measurement rather than the total score of the test being analyzed. The shallow descent of the ANOVA2 method at the low end of the score scale could be due to random guessing by low scoring examinees.

Insert Figure 1 about here

Figure 2 presents additional unsmoothed results from the application of the present method to the Natural Sciences subtest. Conditional alpha coefficient estimates are denoted by Alpha and were computed using (12). Estimated conditional standard errors of measurement in prediction are denoted by CSEMP and are the same values as reported in Figure 1 for the ANOVA2 method. They were computed from (10). Conditional standard error of prediction (CSEP) estimates were computed using (9) and are the present method's estimates of conditional quantities the average of which Lord and Novick (1968, chap. 3) refer to as standard error of prediction (SEP). Conditional true score standard deviations (CTSSD) were computed using (11). The variations of CSEP and CTSSD across the score scale are very similar to each other and fairly similar to the variation in the CSEMP estimates, though

the latter are more smooth and less extreme than both of the former.

Insert Figure 2 about here

Under the approach taken here, the CSEMP estimates are analogs for CSEM estimates. The conditional alpha coefficients assess the relative contributions of CSEMP and CTSSD to CSEP. If the conditional alpha equals ½, then CSEMP equals CTSSD. If the conditional alpha is greater than ½, then CTSSD is greater than CSEMP, and the reverse holds when it is less than ½. Within the present framework, where conditioning is on the observed score and accuracy is interpreted as predicting an examinee's score on retesting, another appropriate index appears to be CSEP which measures all the residual variability in the prediction. For example, CSEP estimates may be used to construct approximate conditional confidence intervals for $\mu(T|X) = \mu(X'|X)$, if two full-length parallel tests have been administered and distributional assumptions are justified for the $X'|X$ conditional distributions. Let N_X denote the number of examinees with score X and k the designated quantile from the appropriate distribution. Then

$$P[\bar{X}'|X) - k(\gamma/2)s(X'|X)/N_X^{1/2} \leq \mu(T|X) \leq (\bar{X}'|X) + k(1 - (\gamma/2)\, s(X'|X)/N_X^{1/2}] = (1 - \gamma).$$

Both forms of the licensure examination yielded similar results, and so data from only one form is presented in Figures 3 and 4 which are analogous to Figures 1 and 2 for Natural Sciences except that an additional method, regression, was applied to the licensure examinations and this is illustrated in Figure 3. In the regression method, a cubic regression was used to smooth

---

Insert Figures 3 and 4 about here

---

the estimates obtained from the difference method as described by Feldt et al. (1985). Neither the regression method nor the binomial method used grouping, but the two ANOVA methods were calculated from grouped data with interval widths of 10 score points. Estimates for the two ANOVA methods are only given for that part of the score scale where there were enough examinees to insure stable estimation with these methods. Estimates for the binomial and regression methods extend down to the lowest observed score.

Estimates from the binomial and ANOVA methods on the licensure examination were quite similar in shape to the Natural Sciences estimates except that the estimates from the two ANOVA methods showed slightly more agreement on the licensure examination as is illustrated in Figure 3. The regression method yielded markedly different results. Its estimates reached a maximum, which surprisingly exceeded the binomial method maximum, somewhat to the right of the score scale midpoint and declined asymmetrically therefrom with a slope reversal at the extreme upper end of the score scale, while the two ANOVA methods followed the shape of the binomial estimates at least over that portion of the score scale which yielded stable estimates for the ANOVA methods. The graphs of the regression estimates were essentially identical for both forms of the licensure examination, and these two forms were composed of substantially different item sets and administered to different groups of examinees. However, there were only about fifty examinees with scores between 120 and 170 and about fifty examinees with scores between 280 and 290 on each form, so some caution may be required in interpreting the regression estimates at the score scale extremes.

It was previously asserted that on average the difference method would overestimate $\sigma^2(E|X)$ but that this overestimation of $\sigma^2(E|X)$ may make it a

better estimate, at least on average, of $\sigma^2(E|T)$, and that the ANOVA1 method would on average underestimate $\sigma^2(E|T)$. Recall the result: $\mu[\sigma^2(E|T)] = \mu[\sigma^2(E'|X)]$. A comparison of the ANOVA1 estimates to the ANOVA2 estimates in Figures 1 and 3 supports the conclusion that the ANOVA1 estimates are on average slightly underestimating $\sigma^2(E|T)$, but a comparison of the regression estimates to the ANOVA2 estimates indicates that the regression method is not only, on average, overestimating $\sigma^2(E|X)$ but also $\sigma^2(E|T)$ at least over that portion of the score scale where stable ANOVA1 and ANOVA2 estimates are available. These conclusions are also supported by the results from the other tests which were analyzed but whose graphs are not presented.

Figure 4 presents estimates obtained from the ANOVA2 method for the licensure examination. As was the case for the Natural Sciences subtest, the CSEP and CTSSD estimates vary similarly across the score scale and are less smooth and more extreme in their variation than are the CSEMP estimates. Explanations for these consistent findings may be found in an examination of the formulas for these statistics. Substituting (10) and (11) into (9) yields

$$s^2(X'|X^*=q) = 4m_2[(MS_p|X^*=q) - (\tfrac{1}{2})(MS_{PM}|X^*=q)]. \qquad (13)$$

The similitude between (13) and (11) explains the concordance in behavior between the CSEP and CTSSD estimates. One possible reason that these estimates are less smooth than the CSEMP estimates is that they depend on the main effect mean square, MS_p, in addition to the interaction mean square, MS_{PM}. MS_p has larger sampling variability than MS_{PM}, and the two combined have even larger sampling variability.

As was mentioned earlier, a direct test of the assumptions underlying the present paper's ANOVA2 method is not readily available unless two full-length parallel tests have been administered. A crude check on the assumptions, not requiring two full-length tests, may be had from a comparison between the observed standard deviation for the total test score, s(X), and the marginal

standard deviation resulting from the model underlying the ANOVA2 method which is given by the square root of $(m[s^2(X'|X^*)] + s^2[m(X'|X^*)])$ where m denotes the mean. This comparison was done for all six of the tests that were analyzed, and good agreement was found between the two standard deviations for all of them. The relative error, which is defined as the difference between the two standard deviations divided by s(X), never exceeded 5%.

Recall that interval widths of 4 score points for the Natural Sciences subtest and 10 score points for the licensure examination were used in applying the ANOVA2 method. If the width of these intervals were decreased, then the CSEP estimates might show a slight decrease also because the CTSSD would decrease, but the CSEMP estimates should remain relatively unchanged. Because CSEP appears to be a useful index and because there is a relatively simple method for obtaining ungrouped CSEP estimates, it is worthwhile to investigate the effect of grouping on CSEP estimates and in the process illustrate this simple method for computing ungrouped estimates.

Let $X_1^* = bY_1 + c$ denote Y_1 linearly equated to X and let $X_2^* = dY_2 + e$ denote Y_2 linearly equated to X. Ungrouped estimates for CSEP may be obtained by regressing X_2^* onto X_1^* and then computing the mean of the squared residuals, $(X_2^* - \hat{X}_2^*)^2$, for each value of X_1^*. This method was applied to the licensure examination using a linear regression of X_2^* onto X_1^*, and the results are presented in Figure 5 wherein an asterisk represents an ungrouped CSEP

Insert Figure 5 about here

---

estimate and the joined squares represent unsmoothed CSEP estimates using the ANOVA2 method with group intervals of 10 score points. It should be noted that an asterisk can be based on anywhere from 1 to 150 examinees with the

smaller sample sizes occurring at the extremes of the score scale. An unweighted cubic polynomial regression of the ungrouped CSEP estimates onto X_i^* is also presented in Figure 5.

There is little evidence for consistent and significant inflation due to grouping for the CSEP estimates presented in Figure 5. The unsmoothed grouped estimates are fairly close to the smoothed ungrouped estimates, though the latter offer some improvement over the former in stability at the extremes of the score distribution. Similar results were found for the other form of the licensure examination. Smoothing the grouped and especially the ungrouped CSEP estimates would likely prove worthwhile in most applications where the sample size is less than enormous. Finally, Davidian and Carroll (1987) present a variety of sophisticated estimation techniques that could prove adaptable to the present ungrouped method and which may improve the method's performance especially with smaller sample sizes.

Summary

The focus of previous research has been on the parameter $\sigma(E|T)$, which is here called conditional standard error of measurement (CSEM), for assessing the accuracy of a test at different values on its score scale. Estimates for $\sigma(E|T)$ have been derived from strong true score models and from weak true score models, but all have problems or limitations. Lord (1984) has discussed the restrictions that apply to CSEM estimates obtained from the binomial error model, and it has been noted that parametric IRT estimates are computationally intensive and require assumptions which may restrict their application in certain situations, e.g., heterogeneous licensure examinations. The present paper has extensively elaborated on criticisms voiced by Lord (1984) concerning weak true score methods for estimating $\sigma(E|T)$. It has shown that the two varieties of weak true score methods: difference methods and ANOVA

methods, use estimates akin to estimates of $\sigma^2(E|X)$ as substitutes for estimates of $\sigma^2(E|T)$. This fact has not always been recognized. It has also shown that difference methods are overestimating $\sigma^2(E|X)$, on average, but that this may make their estimates better estimates of $\sigma^2(E|T)$, at least on average, because $\mu[\sigma^2(E|X)] \leq \mu[\sigma^2(E|T)] = \sigma^2(E)$. Nonetheless, difference methods are conditioning on the wrong quantity, X, and not T so their estimates are approximate at best. It was also shown that ANOVA methods are more accurately estimating $\sigma^2(E|X)$ and as a consequence they are underestimating $\sigma^2(E|T)$, at least on average, but that the average of this underestimation decreases with increases in the reliability of the test. In addition, it was shown that there are other problems with $\sigma^2(E|X)$ as a measure of accuracy at different scores that affect both the difference methods and the ANOVA methods. In particular, E and T are negatively correlated when the conditioning is on X, and $\sigma^2(E|X)$ is constrained to equal $\sigma^2(T|X)$ for all values of X.

Though problems with the estimates obtained from weak true score methods were documented, it was also noted that these methods have several desirable features. They make only weak assumptions; they are conceptually simple; and they are relatively undemanding computationally. This led to the development of a weak true score method based on ANOVA which estimates alternative parameters that can be used in place of $\sigma^2(E|T)$, and which do not have the undesirable properties of $\sigma^2(E|X)$. The method considers the conditional residual variances obtained from regressing a classically parallel measurement, $X'=T+E'$ onto $X=T+E$, and it decomposes these variances into true and error parts. In particular, it was shown that E' and T are uncorrelated given X so that $\sigma^2(X'|X) = \sigma^2(T|X) + \sigma^2(E'|X)$. The conditional residual standard deviations, $\sigma(X'|X)$, are called conditional standard errors of

prediction (CSEP), whereas the conditional standard deviations of just the error parts, $\sigma(E'|X)$, are called conditional standard errors of measurement in prediction (CSEMP). It was demonstrated that $\mu[\sigma^2(E'|X)] = \mu[\sigma^2(E|T)] = \sigma^2(E)$, and that if $\sigma^2(E'|X) = 0$ for all X, then $X' = X = T$. Neither of these properties are true for $\sigma^2(E|X)$.

It was proposed that the ANOVA estimates of CSEMP, which were developed within and which only require scores on two classically parallel half-tests and not two parallel full-tests, be used in conjunction with or in place of the other estimates of either $\sigma(E|T)$ or $\sigma(E|X)$ in assessing variability in the accuracy of tests across their score scales. It was also suggested that estimates of CSEP obtained from the same ANOVA method could prove useful in assessing the constancy of observed scores from one testing to another parallel testing across various score levels since CSEP reflects all variability--that due to both true and error scores. It was demonstrated how CSEP estimates could be used to construct conditional confidence intervals for $\mu(X'|X) = \mu(T|X)$ if two full-length parallel tests were available and appropriate distributional assumptions were justified.

The new ANOVA method as well as several previously developed methods were applied to the four subtests (English, Mathematics, Social Sciences, and Natural Sciences) of the ACT Assessment Program using data from 40,000 examinees on each subtest and to two forms of a licensure examination each of which had been administered to approximately 4,000 examinees. The results from the four subtests were very similar and so only the Natural Sciences subtest results were presented. For the same reason, results from only one of the licensure examinations were presented.

The results indicated that the new ANOVA method works well in practice,

and that the CSEMP estimates which it yields could be used in place of or in conjunction with estimates obtained from other methods. The results also indicated that different methods need not always yield the same results, and that the distribution of raw scores can influence the way in which estimates obtained from weak true score methods vary across the score scale. The latter result was also found by Marco (1988) using parametric IRT methods.

The estimates of CSEP obtained from the new ANOVA method were found to be slightly less stable than the CSEMP estimates, which were quite smooth, especially for the licensure examination where only 4,000 examinees were tested. A simple alternative to the ANOVA method, based on regression, which only produces CSEP estimates was presented and illustrated with the licensure examination data. Smoothing these alternative estimates of CSEP produced more stable estimates of CSEP especially at the extremes of the score scale. Comparison of the new ANOVA method estimates with the alternative regression estimates of CSEP indicated that forming groups of adjacent scores rather than using each distinct score in the application of the ANOVA method did not produce appreciable inflation in the CSEP estimates.

References

Aitchinson, J. (1986). The statistical analysis of compositional data. New York: Chapman & Hall.

Baker, F. D. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models. Applied Psychological Measurement, 11, 111-141.

Blixt, S. L., & Shama, D. D. (1986). An empirical investigation of the standard error of measurement at different ability levels. Educational and Psychological Measurement, 46, 545-550.

Committee of AERA, APA, & NCME to Develop Standards. (1985). Standards for educational and psychological testing. Washington, D.C.: American Psychological Association.

Davidian, M., & Carroll, R. J. (1987). Variance function estimation. Journal of the American Statistical Association, 82, 1079-1091.

Feldt, L. S., Steffen, M., and Gupta, N. C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. Applied Psychological Measurement, 9, 351-361.

Hoyt, C. (1941). Test reliability estimated by the analysis of variance. Psychometrika, 6, 153-160.

Jarjoura, D. (1986). An estimator of examinee-level measurement error variance that considers test form difficulty adjustments. Applied Psychological Measurement, 10, 175-186.

Keats, J. A. (1957). Estimation of error variances of test scores. Psychometrika, 22, 29-41.

Lindley, D. V. (1947). Regression lines and the linear functional relationship. Journal of the Royal Statistical Society, 9, 218-244.

Livingston, S. A. (1982). Estimation of the conditional standard error of measurement for stratified tests. Journal of Educational Measurement, 19, 135-138.

Lord, F. M. (1955). Estimating test reliability. Educational and Psychological Measurement, 15, 325-336.

Lord, F. M. (1984). Standard errors of measurement at different ability levels. Journal of Educational Measurement, 21, 239-243.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Marco, G. L. (1988). Does the use of test assembly procedures proposed in legislation make any difference in test properties and in test performance of blacks and white test takers? Applied Measurement in Education, 1, 109-133.

Mollenkopf, W. G. (1949). Variation of the standard error of measurement. Psychometrika, 14, 189-229.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), Educational Measurement (pp. 560-620). Washington, D.C..: American Council on Education.

Van der Linden, W. J. (1979). Binomial test models and item difficulty. Applied Psychological Measurement, 3, 401-411.

Appendix

First, note that

$$\sigma^2(E) = \mu[\sigma^2(E|T)] + \sigma^2[\mu(E|T)] = \mu[\sigma^2(E|T)], \tag{A1}$$

since $\mu(E|T) = 0$ for all T by assumption. Next, note that

$$\sigma^2(E) = \mu[\sigma^2(E|X)] + \sigma^2[\mu(E|X)]. \tag{A2}$$

The second term on the right may be rewritten as

$$\begin{aligned} \sigma^2[\mu(E|X)] &= \sigma^2[\mu(X - T|X)] \\ &= \sigma^2[\mu(X|X) - \mu(T|X)] \\ &= \sigma^2[X - \mu(T|X)] \end{aligned} \tag{A3}$$

which is greater than zero unless $\mu(T|X) = X$ for all X, a condition not satisfied in general. For instance, if the regression of T on X is linear, then

$$\mu(T|X) = \rho(X,X')X + [1 - \rho(X,X')]\,\mu(X) \tag{A4}$$

which equals X for all X if and only if $\rho(X,X') = 1$. Consequently,

$$\mu[\sigma^2(E|X)] \leq \mu[\sigma^2(E|T)]. \tag{A5}$$

Substituting (A4) into (A3) and then (A3) into (A2) yields:

$$\begin{aligned} \sigma^2(E) &= \mu[\sigma^2(E|X)] + \sigma^2[X - (\rho(X,X')X + (1-\rho(X,X'))\mu(X))] \\ &= \mu[\sigma^2(E|X)] + \sigma^2[(1 - \rho(X,X'))(X - \mu(X))] \\ &= \mu[\sigma^2(E|X)] + [1 - \rho(X,X')]^2\,\sigma^2(X) \\ &= \mu[\sigma^2(E|X)] + [\sigma^2(E)]^2/\sigma^2(X). \end{aligned} \tag{A6}$$

Rewriting the above as

$$\sigma^2(E) - [\sigma^2(E)]^2/\sigma^2(X) = \mu[\sigma^2(E|X)]$$

and factoring $\sigma^2(E)$ out of the left side gives

$$\sigma^2(E)\,\rho(X,X') = \mu[\sigma^2(E|X)]\,. \tag{A7}$$

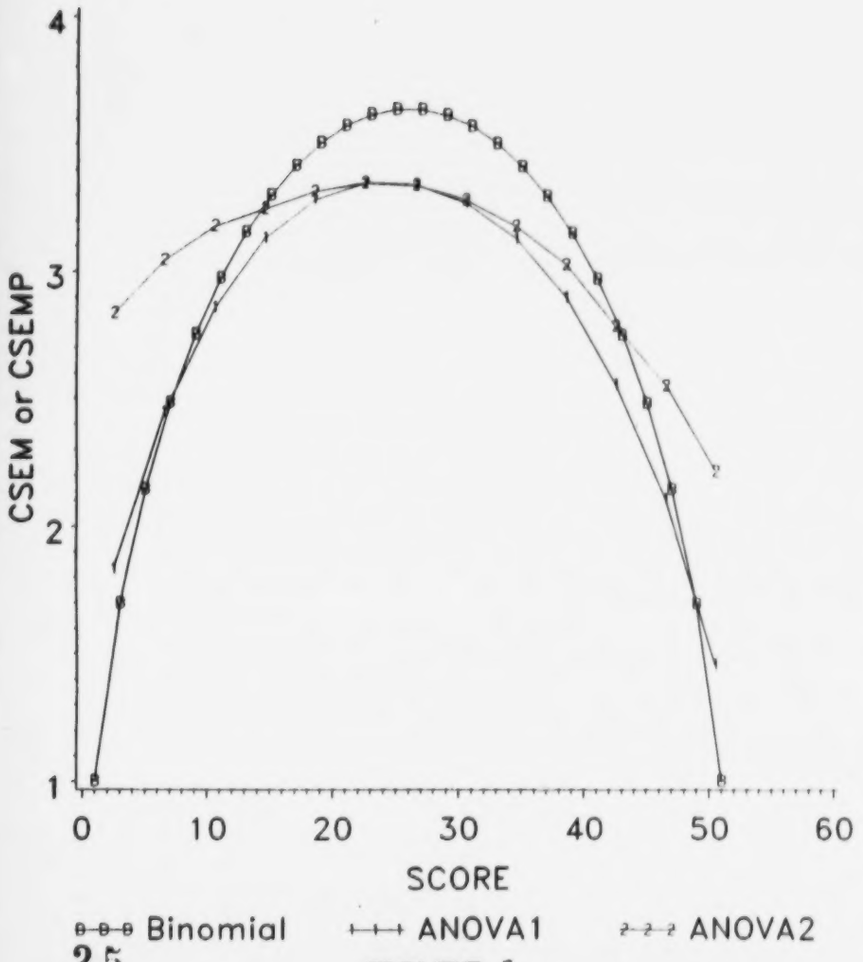


FIGURE 1

Plot of Estimated Conditional Standard Errors of Measurement for the Natural Science Test.

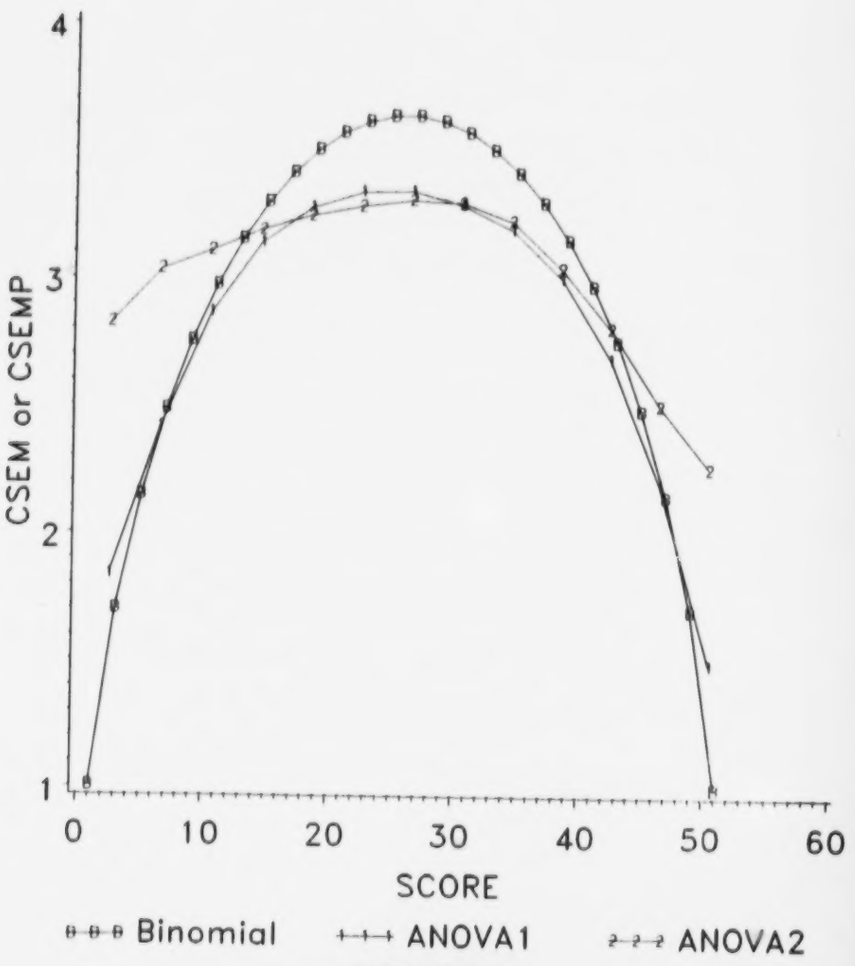


FIGURE 1A

Plot of Estimated Conditional Standard Errors of Measurement for the Social Science Test.

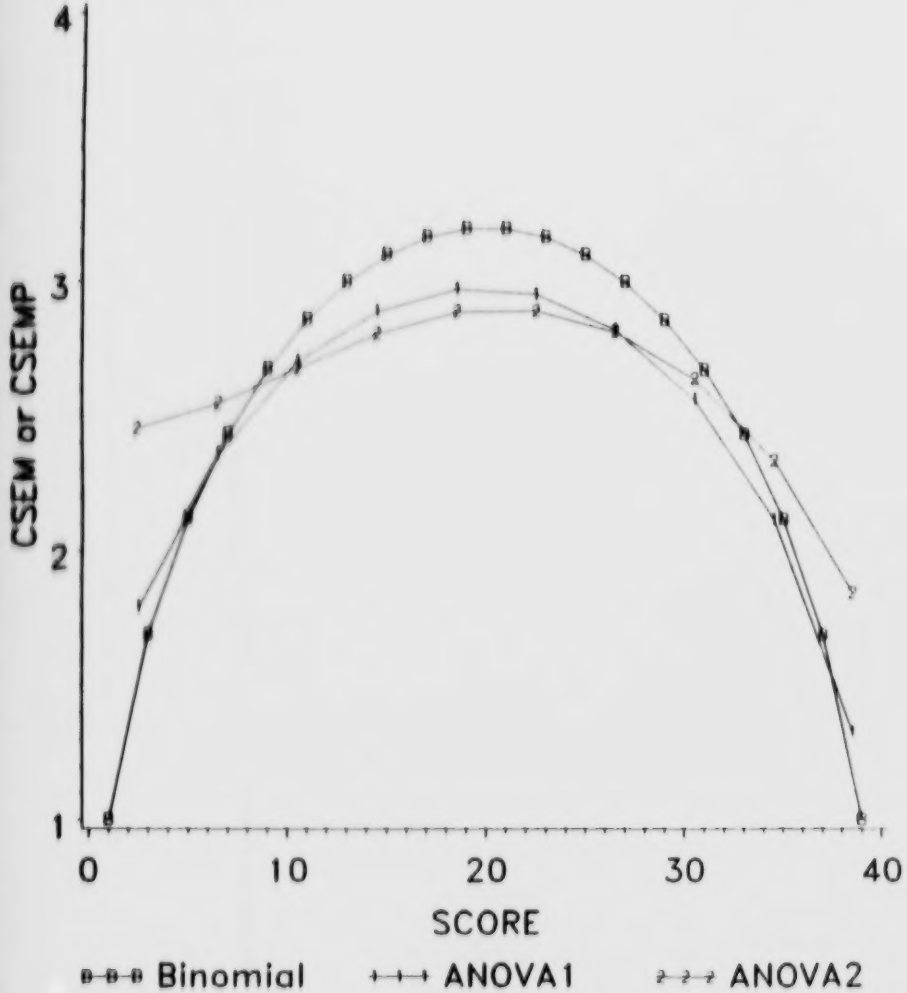


FIGURE 1B

Plot of Estimated Conditional Standard Errors of Measurement for the Mathematics Test

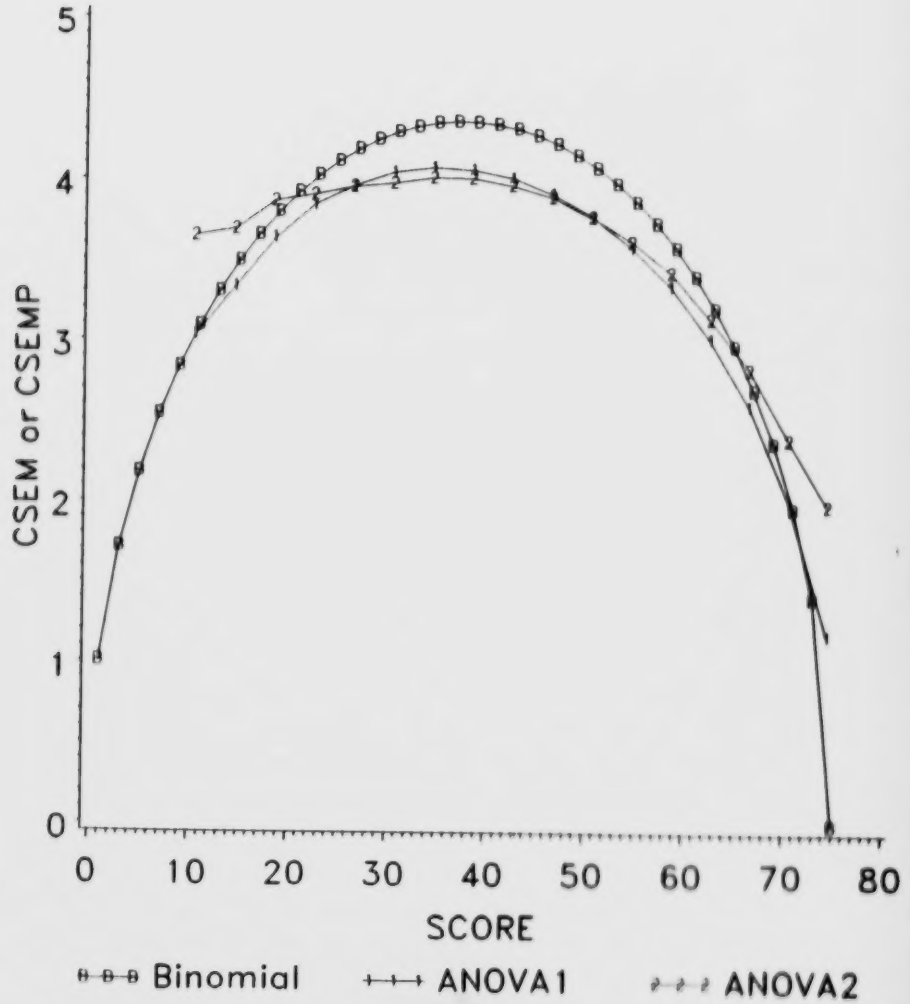


FIGURE 1C

Plot of Estimated Conditional Standard Errors of Measurement for the English Test.

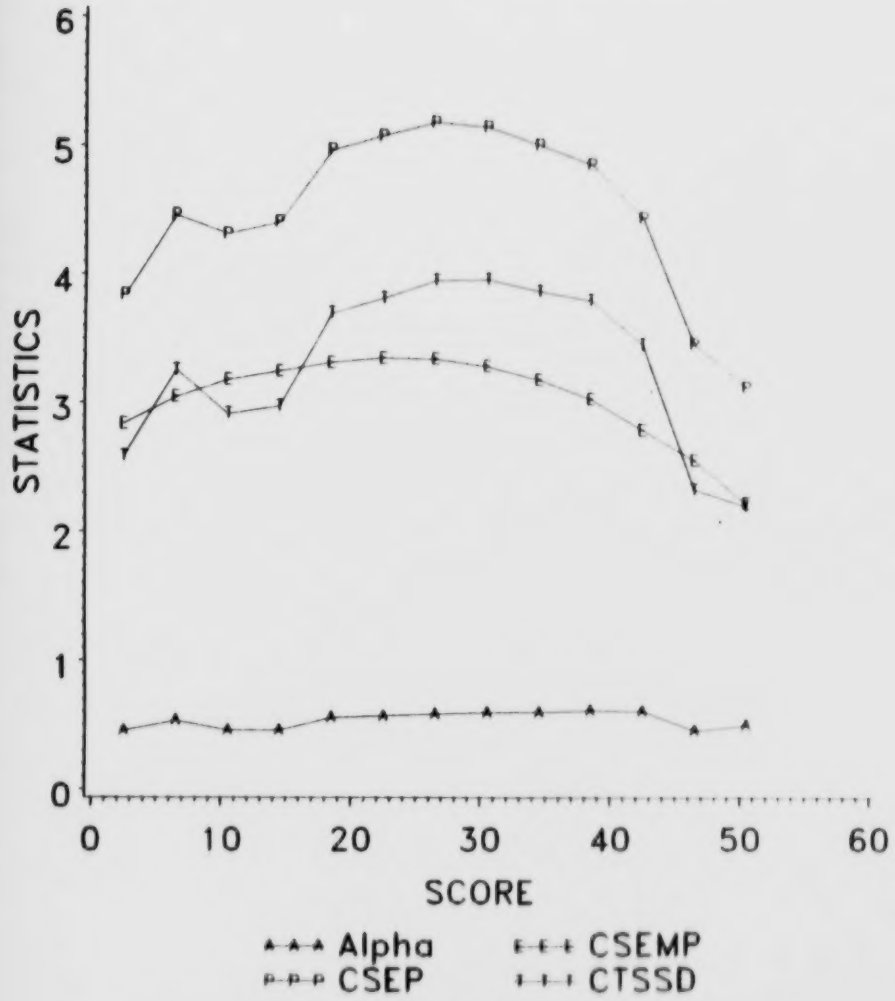


FIGURE 2
Plot of Conditional Statistics from the ANOVA2 Method for the Natural Science Test.

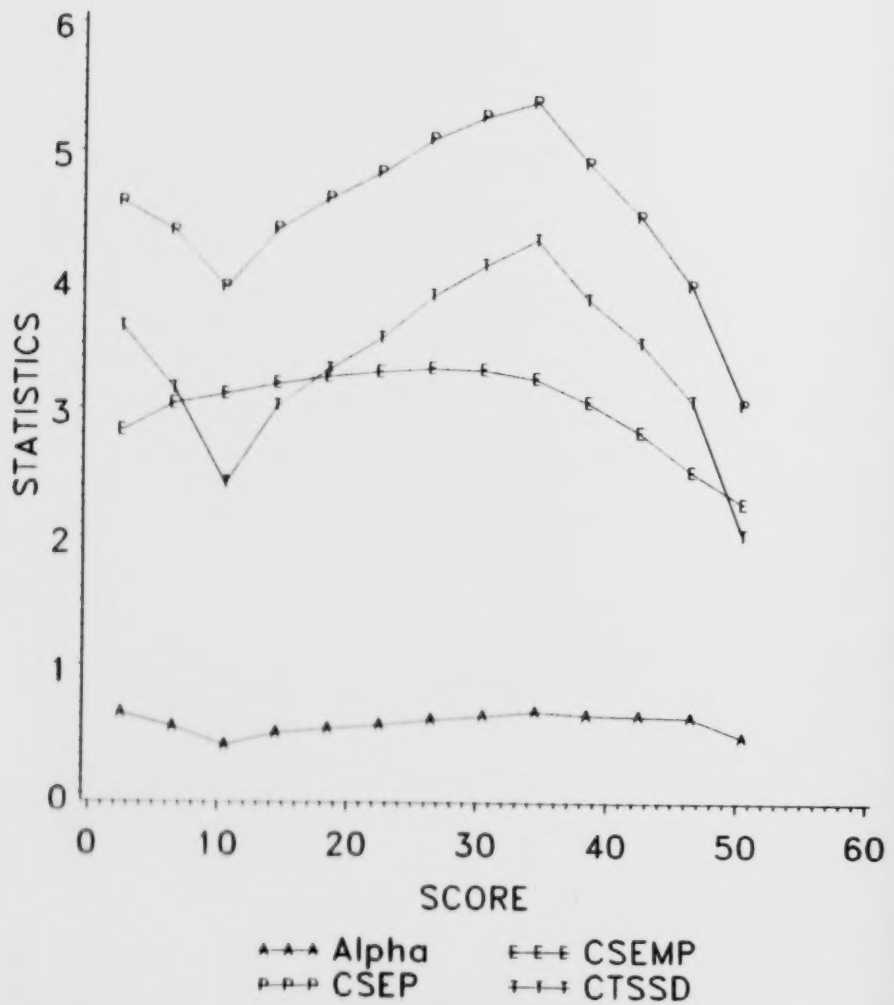


FIGURE 2A
Plot of Conditional Statistics from the ANOVA2 Method for the Social Science Test.

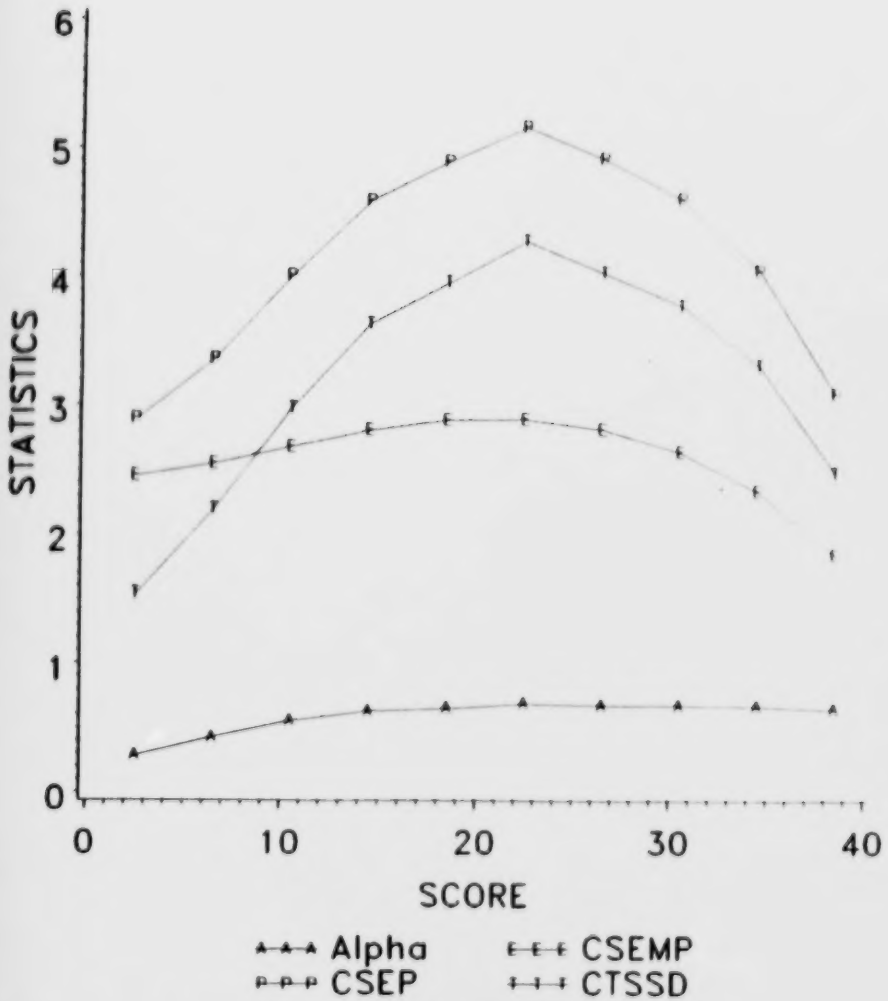


FIGURE 2B

Plot of Conditional Statistics from the ANOVA2 Method for the Mathematics Test.

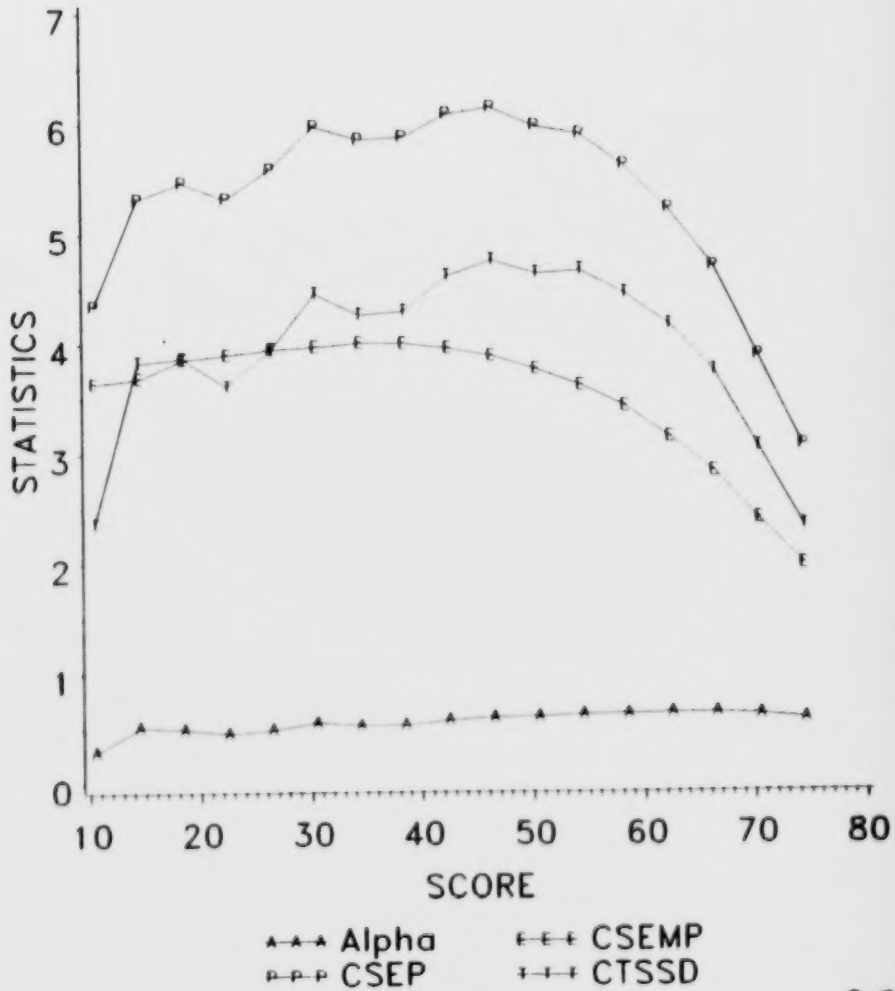
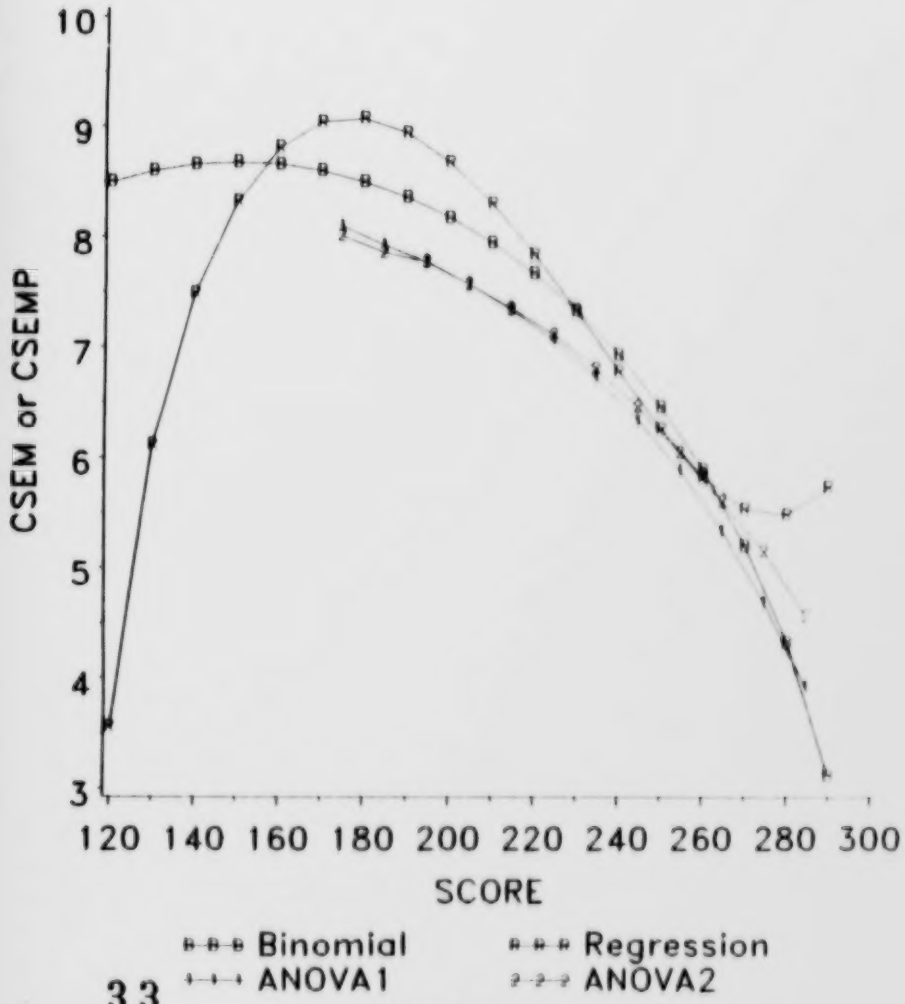


FIGURE 2C

Plot of Conditional Statistics from the ANOVA2 Method for the English Test.



33

FIGURE 3

Plot of Estimated Conditional Standard Errors of Measurement for the 1987 Licensure Examination.

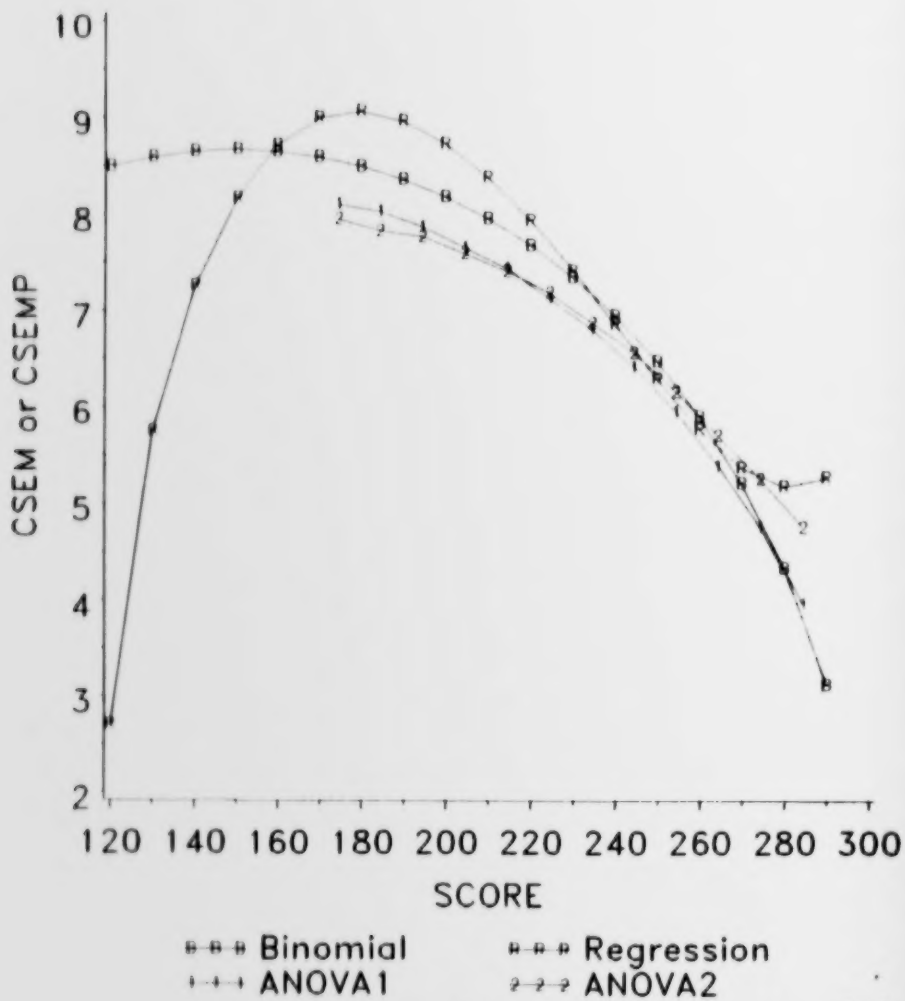


FIGURE 3A

34

Plot of Estimated Conditional Standard Errors of Measurement for the 1986 Licensure Examination.

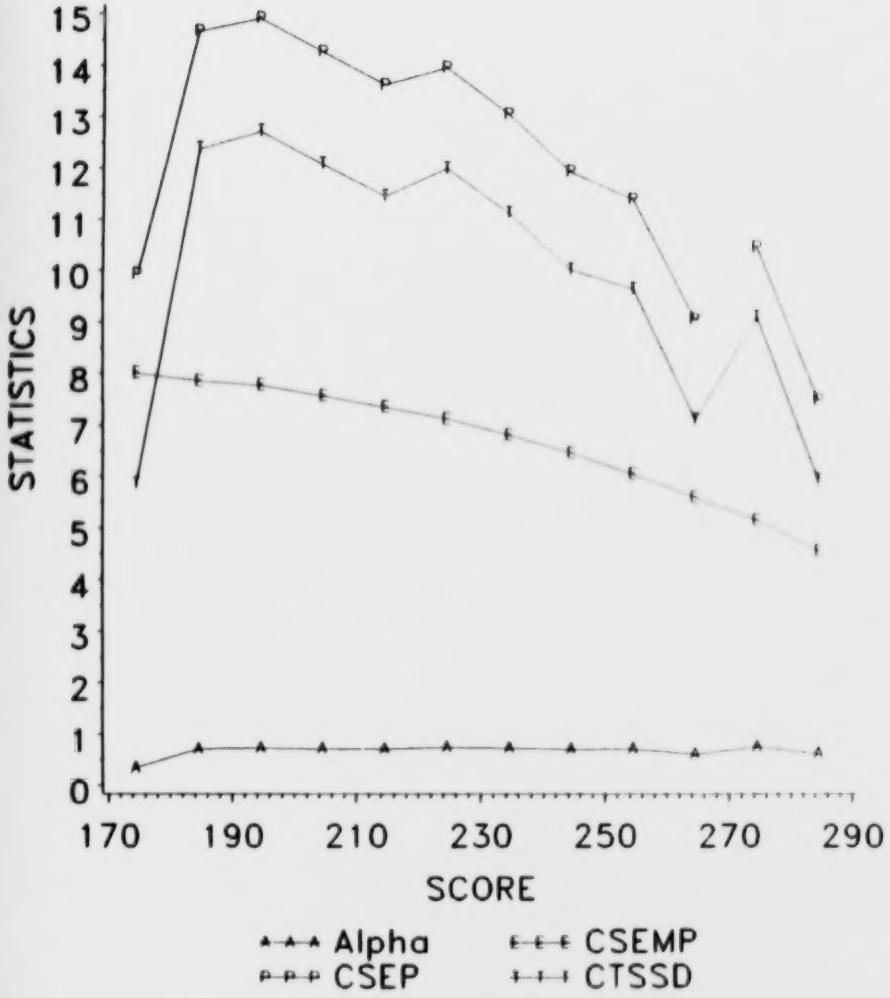


FIGURE 4

Plot of Conditional Statistics from the ANOVA2 Method for the 1987 Licensure Examination.

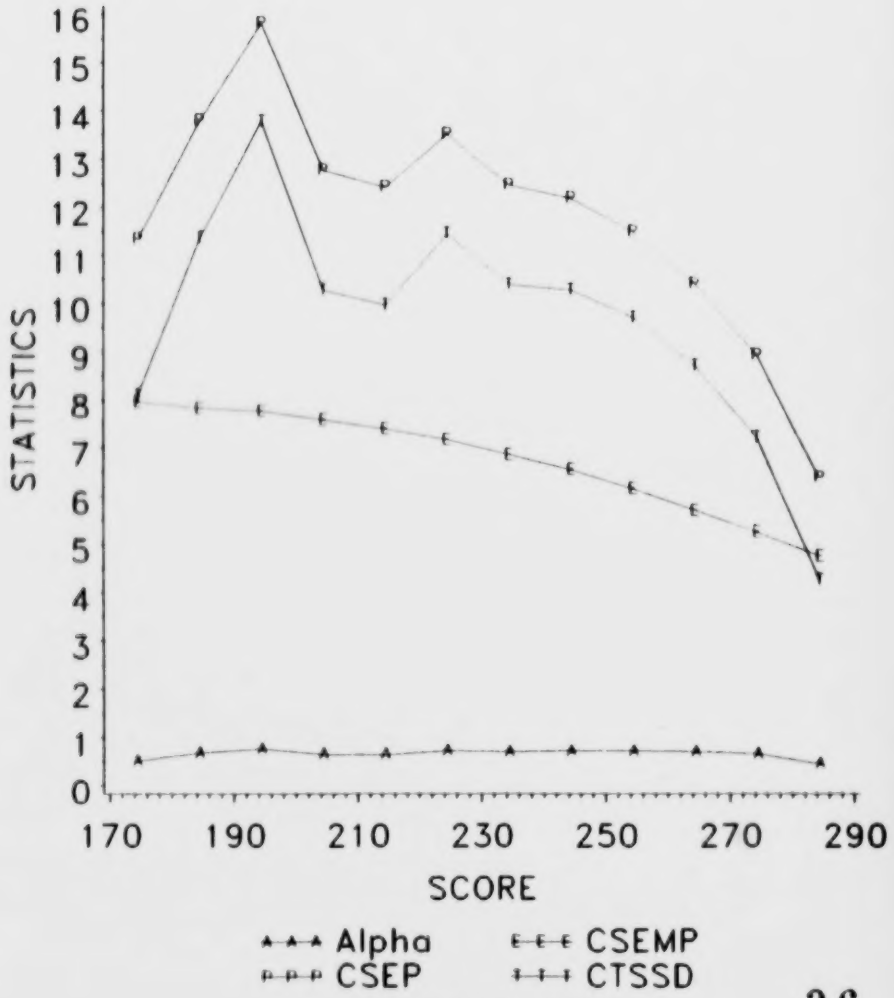


FIGURE 4A

Plot of Conditional Statistics from the ANOVA2 Method for the 1986 Licensure Examination

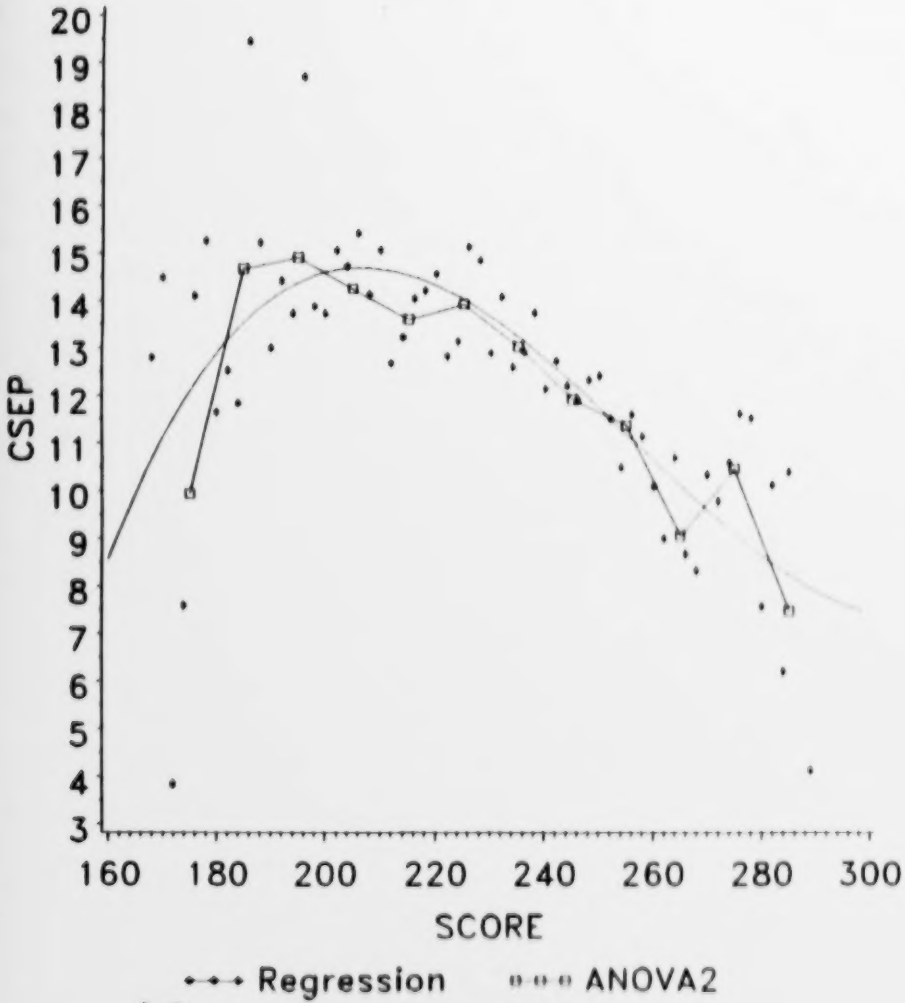


FIGURE 5

Plot of Estimated Conditional Standard Errors of Prediction for the 1987 Licensure Examination.

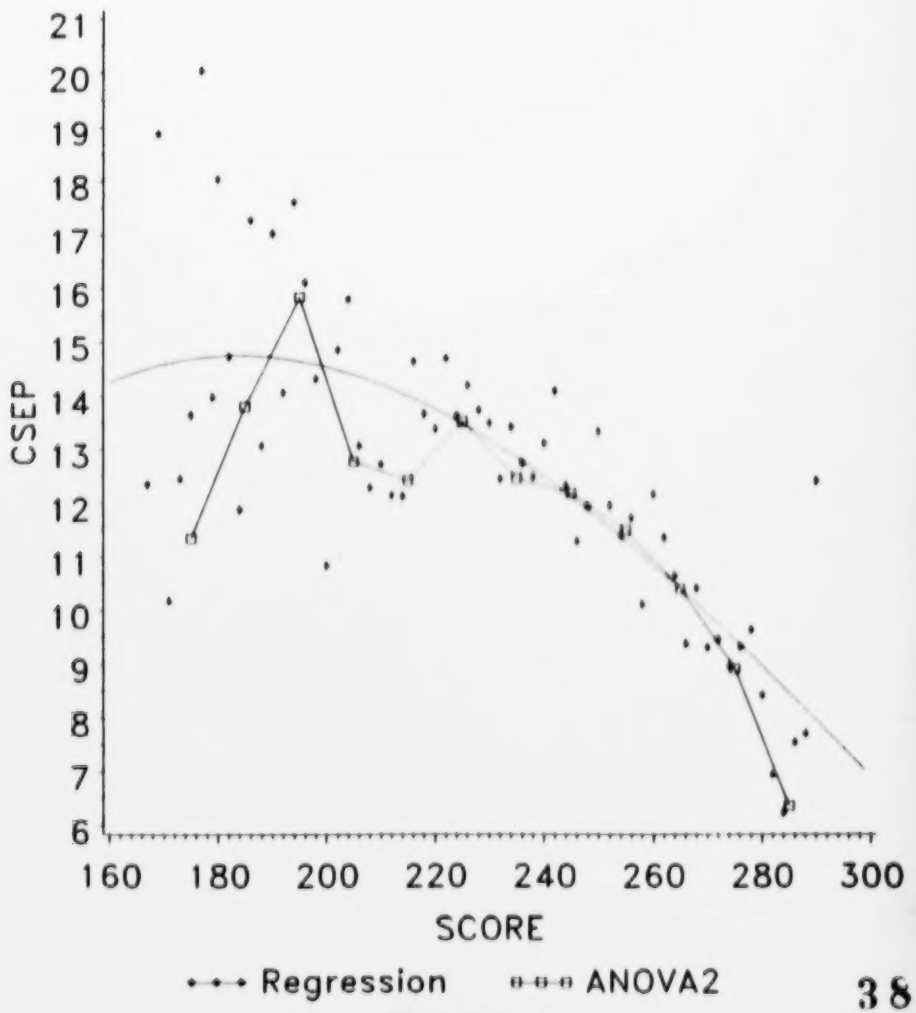


FIGURE 5A

Plot of Estimated Conditional Standard Errors of Prediction for the 1986 Licensure Examination.